DOCUMENT RESUME

ED 435 643 TM 030 243

| | |
|---|---|
| AUTHOR | Chen, Shu-Ying; Ankenmann, Robert D.; Spray, Judith A. |
| TITLE | Exploring the Relationship between Item Exposure Rate and Test Overlap Rate in Computerized Adaptive Testing. |
| INSTITUTION | American Coll. Testing Program, Iowa City, IA. |
| REPORT NO | ACT-RR-99-5 |
| PUB DATE | 1999-09-00 |
| NOTE | 34p. |
| AVAILABLE FROM | ACT Research Report Series, PO box 168, Iowa City, IA 52243-0168. |
| PUB TYPE | Reports - Evaluative (142) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | *Adaptive Testing; *Computer Assisted Testing; Item Banks; *Test Items; Test Length |
| IDENTIFIERS | *Item Exposure (Tests); Overlap Hypothesis |

ABSTRACT

This paper presents a derivation of an average between-test overlap index as a function of the item exposure index, for fixed-length computerized adaptive tests (CAT). This relationship is used to investigate the simultaneous control of item exposure at both the item and test levels. Implications for practice as well as future research are also discussed. An appendix demonstrates that the expected value of the between-test overlap for a fixed-length CAT under completely randomized item selection, is equal to the fixed test length divided by the item pool size. (Contains 5 tables, 2 figures, and 12 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ACT Research Report Series 99-5

ED 435 643

# Exploring the Relationship Between Item Exposure Rate and Test Overlap Rate in Computerized Adaptive Testing

Shu-Ying Chen

Robert D. Ankenmann

Judith A. Spray

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

P. A. Farrant

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM030243

ACT

September 1999

For additional copies write:

ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

© 1999 by ACT, Inc. All rights reserved.

# Exploring the Relationship Between Item Exposure Rate and Test Overlap Rate in Computerized Adaptive Testing

**Shu-Ying Chen and Robert D. Ankenmann**
*The University of Iowa*

**Judith A. Spray**
*ACT, Inc.*

**Abstract**

This paper presents a derivation of an average between-test overlap index as a function of the item exposure index, for fixed-length computerized adaptive tests. This relationship is used to investigate the simultaneous control of item exposure at both the item and test levels. Implications for practice as well as future research are also discussed.

**Exploring the Relationship Between Item Exposure Rate and Test Overlap Rate in Computerized Adaptive Testing**

The popularity of computerized adaptive tests (CATs) has increased in recent years due to the significant progress of computer technology. Many conventional paper-and-pencil (P&P) tests, like the Graduate Record Examination (GRE) and the Armed Services Vocational Aptitude Battery (ASVAB), are now offered in a CAT format. One practical advantage of CATs is that they can be administered on a flexible schedule rather than at fixed times. The convenience and flexibility for examinees, however, may severely compromise test security if item exposure is not well controlled. Because test security is always an important concern, especially in high stakes testing programs (e.g., college admissions, or certification and licensure), CATs cannot be implemented effectively in practice unless item exposure is well controlled.

Way (1998) stated that, to date, the methods used to avoid item overexposure in CATs fall into two general categories: (a) randomized item selection (e.g., McBride & Martin, 1983; Bergstrom, Lunz, & Gershon, 1992; Way Zara, & Leahy, 1996); and (b) conditional item selection (e.g., Sympson & Hetter, 1985; Davey & Parshall, 1995; Stocking & Lewis, 1995, 1998). Regardless of the item exposure control method used, item exposure rate and average item overlap are two indices commonly used to track item exposure in CATs (Way, 1998). Item exposure rate refers to the relative frequency with which an item is presented across all CAT administrations, that is, the proportion of all CATs in which an item is administered. Average item overlap is defined by Way

(1998) as the proportion (or percentage) of items shared by pairs of exams, averaged across all possible pairwise comparisons. It is important to note that Mills and Stocking (1996) use the term *item overlap* in referring to "the extent to which one item may cue the correct response to another item or the extent to which two items depend on the same specific knowledge" (p. 294). To avoid confusion, and to provide a more accurate and descriptive nomenclature, we introduce the following terminology and definitions: (a) For a pairwise comparison between two fixed-length CATs that have been administered, the *between-test overlap* is the proportion of items on one test that also appear on the other test (i.e., the proportion of shared items); and (b) the *average between-test overlap* is the arithmetic mean of the between-test overlaps across all possible pairwise comparisons. Furthermore, we use the terms *average between-test overlap* and *test overlap rate* interchangeably. The average between-test overlap, as defined above, is equivalent to the average item overlap defined by Way (1998). By considering both the item exposure rate and the average between-test overlap, item exposure can be monitored at the individual item level as well as the test level.

Despite the importance of both item exposure rate and test overlap rate in tracking item exposure control, few studies have investigated the effects of simultaneously controlling the magnitudes of these two indices. While most research to date has focused on item exposure control at the individual item level, Davey and Parshall (1995) proposed a conditioned item exposure control method designed to function at both the item and test levels. Although this method reduces the amount of test overlap and is more general than methods that function only at the individual item

level, it fails to control the test overlap rate exactly, that is, it fails to ensure desired levels of test security. Research with more comprehensive methods of controlling the item exposure rate and the test overlap rate simultaneously may be useful.

Based on the conceptual definitions of item exposure rate and average between-test overlap, it is to be expected that these two indices are highly related. If the average between-test overlap could be expressed as a function of item exposure rates, then it would not be necessary to undertake time-consuming pairwise comparisons of CATs to determine all between-test overlaps. Rather, the average between-test overlap could be more simply computed once the item exposure rates were known. Such a simplification would be especially efficient when the number of CATs administered (hence the number of pairwise comparisons) is large. Furthermore, if the average between-test overlap could be expressed as a function of the item exposure rates, then the relationship between these two indices could be investigated directly and easily. This, in turn, may provide insights into CAT design and implementation considerations relevant to the simultaneous control of item exposure rates and average between-test overlap.

The purpose of this paper is to present an analytical derivation for the mathematical form of an average between-test overlap index as a function of the item exposure index, for fixed-length CATs. This algebraic relationship is used to investigate the simultaneous control of item exposure at both the item and test levels. Implications for practice as well as future research are also discussed.

## Theoretical Background

To facilitate the mathematical derivations which follow, consider the following hypothetical example: An item pool consists of $n = 10$ items, from which $p = 4$ fixed-length CATs are administered, each CAT consisting of $k = 5$ items. Case 1 in Table 1 shows the items that were administered in each of the four CATs, and Table 2 shows the number of times each item was used ($m_i$).

---

See Tables 1-2 at end of report.

---

Making a pairwise comparison of the items administered in $p_1$ and $p_2$ (i.e., the first and second CATs, respectively), two items (2 and 4) were administered in both CATs. Thus, the between-test overlap for the $p_1 p_2$ comparison is 2/5. Calculating the between-test overlap for each possible pairwise comparison and averaging across all six such comparisons yields an average between-test overlap of

$$\bar{T} = \frac{2/5 + 4/5 + 3/5 + 1/5 + 3/5 + 3/5}{6}, \quad (1)$$

which can be written as

$$\bar{T} = \frac{2+4+3+1+3+3}{5(6)}. \quad (2)$$

*Algebraic Form of the Average Between-Test Overlap*

The numerator in Equation 2 is equivalent to the total number of times items were shared between pairs of CATs, across all possible pairwise comparisons. In general, this total is mathematically determined by

$$\sum_{i=1}^{n}\binom{m_i}{2}. \tag{3}$$

Thus, in general, the average between-test overlap is mathematically defined as

$$\overline{T} = \frac{\sum_{i=1}^{n}\binom{m_i}{2}}{k\binom{p}{2}} = \frac{\sum_{i=1}^{n} m_i(m_i-1)}{kp(p-1)}; \tag{4}$$

where $p$ denotes the number of fixed-length CATs administered, $k$ denotes the number of items in each of the CATs, $n$ denotes the number of items in the pool, and $m_i$ denotes the number of times item $i$ was administered across all $p$ CATs.

The item exposure rate of item $i$ (i.e., the proportion of all CATs in which an item is administered) is defined as

$$r_i = \frac{m_i}{p}, \quad i = 1, 2, 3, \ldots, n. \tag{5}$$

Based on this definition, the sum of the item exposure rates across all items in the pool must equal the fixed test length, that is,

$$\sum_{i=1}^{n} r_i = k. \tag{6}$$

Thus, for any given fixed test length and pool size, the average item exposure rate will always be a constant:

$$\bar{r} = \frac{\sum_{i=1}^{n} r_i}{n} = \frac{k}{n}. \quad (7)$$

Note that from Equation 7 it is clear that for a given fixed test length and pool size, the average item exposure rate is fixed. In other words, for a given ratio of pool size to fixed test length, the average item exposure rate is fixed, regardless of the number of CATs administered or the quality of the items in the pool.

Dividing both the numerator and denominator of Equation 4 by $p^2$, substituting Equation 5 into Equation 4, and simplifying yields

$$\bar{T} = \frac{\sum_{i=1}^{n} r_i(pr_i - 1)}{k(p-1)}. \quad (8)$$

Expanding the numerator in Equation 8 yields

$$\bar{T} = \frac{p\sum_{i=1}^{n} r_i^2 - \sum_{i=1}^{n} r_i}{k(p-1)}. \quad (9)$$

Substituting Equation 6 into Equation 9 and simplifying yields

$$\bar{T} = \frac{p\sum_{i=1}^{n} r_i^2}{k(p-1)} - \frac{1}{p-1}. \quad (10)$$

Thus, the average between-test overlap can be expressed as a function of the item exposure rates, fixed test length, and number of CATs administered. Note that

substituting values from Table 2 into Equations 5 and 10 yields a numerical result (i.e., 16/30) identical to that which was obtained from Equation 1, as expected.

*Large-Sample Estimate of the Average Between-Test Overlap*

As the number of CATs that are administered increases (i.e., as $p$ increases)

$$\frac{p\sum_{i=1}^{n} r_i^2}{k(p-1)} - \frac{1}{p-1} \xrightarrow{p\to\infty} \frac{\sum_{i=1}^{n} r_i^2}{k} = \hat{\bar{T}}; \tag{11}$$

thus, we can think of $\hat{\bar{T}}$ as a large-sample estimate of $\bar{T}$. Note that in cases where an item pool can be partitioned into mutually exclusive content area sub-domains, $\sum r^2$ can be partitioned also, with respect to those sub-domains. Thus, it is possible to determine the proportion of the average between-test overlap accounted for by each content area sub-domain.

By completing the square on $r^2$ in the numerator of Equation 11, we obtain

$$\hat{\bar{T}} = \frac{\sum_{i=1}^{n}\left[\left(r-\frac{k}{n}\right)^2 + 2r\frac{k}{n} - \frac{k^2}{n^2}\right]}{k}. \tag{12}$$

Distributing the summation operator over the terms in the numerator of Equation 12, replacing $\sum r$ with $k$ (Equation 6), simplifying, and dividing both the numerator and denominator by $n$ yields

$$\hat{\bar{T}} = \frac{\sum_{i=1}^{n}\left(r_i-\frac{k}{n}\right)^2 + \frac{k^2}{n}}{k} = \frac{\frac{\sum_{i=1}^{n}\left(r_i-\frac{k}{n}\right)^2}{n} + \left(\frac{k}{n}\right)^2}{\frac{k}{n}}. \tag{13}$$

Replacing $k/n$ in Equation 13 with $\bar{r}$ (Equation 7) yields

$$\hat{\bar{T}} = \frac{\frac{\sum_{i=1}^{n}(r_i - \bar{r})^2}{n} + \bar{r}^2}{\bar{r}} = \frac{S_r^2 + \bar{r}^2}{\bar{r}}, \quad (14)$$

where $S_r^2$ denotes the variance of the item exposure rates. Thus, the large-sample estimate of the average between-test overlap is a function of the sample mean and variance of the item exposure rates.

The error in the estimate, $\hat{\bar{T}}$, is defined by

$$\varepsilon(\hat{\bar{T}}) = \bar{T} - \hat{\bar{T}}. \quad (15)$$

Substituting Equation 10 for $\bar{T}$ yields

$$\varepsilon(\hat{\bar{T}}) = \frac{p\sum r^2}{k(p-1)} - \frac{1}{p-1} - \hat{\bar{T}}. \quad (16)$$

Replacing $\frac{\sum r^2}{k}$ with $\hat{\bar{T}}$ (Equation 11) and factoring $\frac{1}{p-1}$ out of every term on the right side of Equation 16 yields

$$\varepsilon(\hat{\bar{T}}) = \frac{1}{p-1}\left[p\hat{\bar{T}} - 1 - (p-1)\hat{\bar{T}}\right], \quad (17)$$

which simplifies to

$$\varepsilon(\hat{\bar{T}}) = \frac{\hat{\bar{T}} - 1}{p-1}. \quad (18)$$

By definition, $0 \le r \le 1$; therefore, $r^2 \le r$, $0 < \sum r^2 / \sum r \le 1$ (i.e., $0 < \sum r^2 / k \le 1$), and $0 < \hat{\bar{T}} \le 1$. Thus, for any $p > 1$, $\varepsilon(\hat{\bar{T}}) \le 0$ and $\hat{\bar{T}} \ge \bar{T}$. That the large-sample estimate of the average between-test overlap consistently exceeds its true value is a desirable

feature: One is guaranteed that the true value of the average between-test overlap is no greater than its large-sample estimate. Another desirable feature of $\hat{\bar{T}}$ is that $\varepsilon(\hat{\bar{T}})$ decreases as $p$ (i.e., the number of CATs administered) increases.

### Implications of the Theory for Practice and Research

Closer inspection of the large-sample estimate of the average between-test overlap ($\hat{\bar{T}}$; Equation 14) reveals that for a given pool size to fixed test length ratio, $\hat{\bar{T}}$ is a linear function of $S_r^2$, the variance of the item exposure rates:

$$\hat{\bar{T}} = \frac{1}{\bar{r}} \cdot S_r^2 + \bar{r}. \tag{19}$$

The slope of the straight line represented by Equation 19 is given by the pool size to fixed test length ratio (i.e., $\frac{1}{\bar{r}} = \frac{n}{k}$), and the $y$-intercept is the reciprocal of this ratio (i.e., $\bar{r} = \frac{k}{n}$). Figure 1 illustrates a family of curves of $\hat{\bar{T}}$ versus $S_r^2$ (defined by Equation 19) for several pool size to fixed test length ratios. Careful examination of Equation 19 and Figure 1 leads to several useful, practical implications for the design and development of CATs.

---

See Figure 1 at end of report.

---

Perhaps the most noteworthy feature of Equation 19 and Figure 1 concerns the $y$-intercept, which is obtained when $S_r^2 = 0$. When there is no variability in the item

exposure rates (as would happen if item selection were completely randomized), the large-sample estimate of the average between-test overlap achieves its lowest possible value, because $\bar{r} > 0$ and $S_r^2 \geq 0$. Thus, $\bar{r} = k/n$ is a lower bound for $\hat{\bar{T}}$. Indeed, under completely randomized item selection for fixed-length CATs, the expected value of the between-test overlap is equal to $k/n$ (see the appendix). The practical implication of this for CAT design is that item pool size must be at least 6.7 times as large as the fixed test length if the average between-test overlap is not to exceed 15%, as prescribed by Way (1998) for CATs used in college admissions decisions. Limiting the test overlap rate to 10% would require a pool size at least 10 times as large as the fixed test length. However, these ratios are deceiving, because they represent minimal prescriptions based on an assumption of completely randomized item selection. In practice, the psychometric qualities of items feature strongly in CAT item selection via some form of an item information function (e.g., Fisher item information), and under these circumstances $S_r^2 > 0$.

Figure 1 clearly indicates that for pool sizes at least 10 times as large as the fixed test length, an item exposure rate variance less than .005 guarantees that the average between-test overlap will be under 15%. An average between-test overlap below 10% is guaranteed when the variance of the item exposure rates is less than .002 with pool sizes at least 14 times as large as the fixed test length (or less than roughly .0014 with pool sizes at least 12 times as large as the fixed test length). More generally, Figure 1 demonstrates that both the pool size to fixed test length ratio (the reciprocal of which is equal to the average item exposure rate) and the variance of the item exposure rates are

important in determining the average between-test overlap. It is important to note that, in addition to Way's (1998) prescriptions for average item exposure and average percent overlap, the variance of the item exposure rates is a crucial element in CAT design. In fact, any given pool size to fixed test length ratio fixes the average item exposure rate, thus necessitating some degree of control over the variance of the item exposure rates if the average between-test overlap is to be controlled. Increasing the pool size to fixed test length ratio, alone, does not guarantee that the average between-test overlap will be maintained within desired limits; the variance of the item exposure rates must be controlled also.

These theoretically based insights are confirmed with empirical results. Table 3 provides a summary of empirical item exposure data obtained from a CAT simulation study conducted by Chen and Ankenmann (1999), and corresponding algebraically obtained between-test overlap data. The item pool consisted of 360 ACT-Math items representing six content area sub-domains (pre-algebra, 23.33%; elementary algebra, 16.67%; intermediate algebra, 15%; coordinate geometry, 15%; plane geometry, 23.33%; and trigonometry, 6.67%) and the fixed test length was 20 items. Thus, the pool size to fixed test length ratio was 18:1. Adaptive item selection was implemented with the Fisher item information function and three parameter logistic item response model, content balancing was implemented with a multinomial model, and the Sympson and Hetter (1985) item exposure control method was used with a desired maximum item exposure rate of .20. A total of 7,000 CATs were simulated, 1,000 at each of seven proficiency levels ($\theta$ = -3, -2, -1, 0, 1, 2, 3). The empirically obtained mean ($\bar{r}$) and

variance ($S_r^2$) of the item exposure rates were used in Equations 10, 14, and 18 to obtain $\bar{T}$, $\hat{\bar{T}}$, and $\varepsilon(\hat{\bar{T}})$, respectively.

---

See Table 3 at end of report.

---

Notice that without content balancing or item exposure control, the variance of the item exposure rates exceeded .01 and the average between-test overlap was about 30%. Content balancing, alone, did little to reduce the variance of the item exposure rates and the average between-test overlap. Item exposure control (in addition to content balancing) was required to reduce the variance of the item exposure rates to .002, which was enough to bring the average between-test overlap under 10%. With a pool size to fixed test length ratio of 18:1, Equation 14 indicates that an item exposure rate variance less than .0025 guarantees an average between-test overlap less than 10%. Notice that across all of the studied conditions the average item exposure rate remained constant. The impact of each studied condition on the variability of the item exposure rates is portrayed more vividly, perhaps, in the frequency distributions shown in Figure 2. Note that the means of these distributions are all the same (i.e., $\bar{r} = \frac{k}{n} = \frac{20}{360} = .0556$).

---

See Figure 2 at end of report.

---

Further empirical confirmation of the significant relationship between the variance of the item exposure rates and the average between-test overlap, under varying item exposure control methods and item pool sizes, is provided in Table 4. All data reported in Table 4 were empirically obtained in a simulation study of CAT item exposure control methods by Chang (1998). In particular, the mean ($\bar{T}$) and variance ($S_T^2$) of the between-test overlaps were empirically obtained from all possible pairwise comparisons of the simulated CATs. Across item exposure control methods and item pool sizes, average between-test overlaps less than 15% were obtained only when the variance of the item exposure rates was less than .003. Under the condition in which no item exposure control was implemented, doubling the item pool size from 360 to 720 (while keeping the fixed test length unchanged at 30) yielded a negligible decrease in the average between-test overlap from 37% to 34%. Changing the pool size to fixed test length ratio, alone, did little to reduce the test overlap rate. This negligible effect was also observed under the McBride and Martin (1983) item exposure control method.

---

See Table 4 at end of report.

---

In the context of high stakes decisions (e.g., college admissions, or certification and licensure), it may be desirable to control not only the average between-test overlap—as prescribed by Way (1998)—but also the maximum value or the variance of the between-test overlaps. Thus, it would be advantageous to determine the functional relationship between the item exposure rates and the variance of the between-test

overlaps, if it indeed exists. Considering the data in Table 4, there is a very strong linear relationship between the variance of the item exposure rates ($S_r^2$) and the variance of the between-test overlaps ($S_I^2$). Across the conditions in which $n = 360$, 99.995% of the variability in $S_I^2$ was accounted for by variability in $S_r^2$. Across conditions in which the pool size was doubled (i.e., $n = 720$), 99.821% of the variability in $S_I^2$ was accounted for by variability in $S_r^2$. However, a consideration of the cases presented in Table 1 suggests that the variance of the between-test overlaps cannot be expressed as a function of the item exposure rates.

Table 5 presents a summary of between-test overlap data corresponding to both examples (cases) presented in Table 1. Both of these cases give rise to identical item exposure rates, because the number of times each item was administered is the same. Note that the only difference between Case 1 and Case 2 concerns the second item administered in $p_1$ and $p_2$, denoted in Table 1 by the circled item numbers. Such a difference has no effect on the average between-test overlap, because the total number of times items are shared in all possible pairwise comparisons remains unchanged; that is, $\overline{T}_{(\text{Case 1})} = \overline{T}_{(\text{Case 2})} = 16/30$. Such a difference does, however, affect the variance of the between-test overlaps; in particular, with $\overline{T} = 16/30$, $S^2_{T(\text{Case 1})} = .0356$ whereas $S^2_{T(\text{Case 2})} = .0222$. The two cases in Table 1 give rise to identical item exposure rates but different between-test overlap variances. In other words, the variance of the between-test overlaps is not uniquely determined by the item exposure rate, number of CATs

administered, and fixed test length. This signals a need for further research concerning the relationship between item exposure rates and between-test overlap variance.

---

See Table 5 at end of report.

---

The mean and variance of the between-test overlap may not be the only important CAT design considerations. The maximum between-test overlap, or perhaps even several percentiles (e.g., $P_{95}$, $P_{90}$, $P_{75}$) of the between-test overlap distribution, may be important also. Therefore, a useful future direction for research might be to investigate the relationship between the distributions of item exposure rate and between-test overlap. Finally, the derivations presented in this paper pertain only to fixed-length CATs, so a natural extension to this work would be to investigate the relationship between item exposure and test overlap in variable-length CATs.

The derivations presented in this paper provide theoretical evidence that, in fixed-length CATs, control of the average between-test overlap is achieved via the sample mean and variance of the item exposure rates. The mean of the item exposure rates is equal to the fixed test length divided by the item pool size, and is therefore easily manipulated. Control over the variance of the item exposure rates can be achieved via the maximum item exposure rate ($r_{max}$). Therefore, this paper also establishes a theoretical basis for concluding that item exposure control methods which implement a specification of $r_{max}$ (e.g., Sympson & Hetter, 1985; Davey & Parshall, 1995; Stocking & Lewis, 1995, 1998) provide the most direct control over the average

between-test overlap. Empirical evidence of this conclusion has been provided by Chang (1998).

## References


Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education, 5*, 137-149.

Chang, S. (1998). *A comparative study of item exposure control methods in computerized adaptive testing*. Unpublished doctoral dissertation, The University of Iowa, Iowa City.

Chen, S., & Ankenmann, R. D. (1999, April). *Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing*. Paper presented at the Annual Meeting of the American Educational Research Association, Montréal, PQ, Canada.

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

DeGroot, M. H. (1986). *Probability and statistics* (2nd ed.). Addison-Wesley, Reading, MA.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 223-236). New York, NY: Academic Press.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9*, 287-304.

Stocking M. L., & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing* (ETS Research Report RR-95-25). Princeton, NJ: Educational Testing Service.

Stocking M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57-75.

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*, 17-27.

Way, W., Zara, A., & Leahy, J. (1996, April). *Modifying the NCLEX™ CAT item selection algorithm to improve item exposure.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

## Appendix

The purpose of this appendix is to show that the expected value of the between-test overlap, for a fixed-length CAT under completely randomized item selection, is equal to the fixed test length divided by the item pool size.

Consider an item pool consisting of $n$ items, from which fixed-length CATs are administered, each CAT consisting of $k$ items. Furthermore, each CAT is formed by randomly selecting $k$ items from the pool. For a pairwise comparison between any two fixed-length CATs that have been administered, let the random variable $Y$ denote the number of items on one test that also appear on the other test (i.e., the number of shared items). Then, the between-test overlap for such a pairwise comparison is given by $Y/k$. Possible values for the random variable $Y$ are $y = 0, 1, 2, \ldots, k$ where $y = 0$ implies no shared items between CATs and $y = k$ implies complete overlap or identical CATs.

Because $Y$ is distributed as a hypergeometric random variable, its probability function is defined as

$$\Pr(Y = y) = \frac{\binom{k}{y}\binom{n-k}{k-y}}{\binom{n}{k}}, \qquad \text{(A-1)}$$

and represents the probability that $Y$, the number of shared items, is equal to $y = 0, 1, 2, \ldots, k$. The expected value of $Y$ (DeGroot, 1986) is

$$\mathrm{E}[Y] = \sum_{y=0}^{k} \frac{\binom{k}{y}\binom{n-k}{k-y}}{\binom{n}{k}} \cdot y = \frac{k^2}{n}. \qquad \text{(A-2)}$$

Therefore, the expected value of the between-test overlap for a fixed-length CAT under completely randomized item selection is

$$\mathrm{E}\left[\frac{Y}{k}\right] = \frac{\mathrm{E}[Y]}{k} = \frac{k}{n}. \tag{A-3}$$

Note, also, that

$$\mathrm{Var}[Y] = \frac{k^2(n-k)^2}{n^2(n-1)}; \tag{A-4}$$

therefore,

$$\mathrm{Var}\left[\frac{Y}{k}\right] = \frac{\mathrm{Var}[Y]}{k^2} = \frac{(n-k)^2}{n^2(n-1)}. \tag{A-5}$$

The expected values defined by Equations A-2 and A-3 can be generalized to account for situations in which the item pool is partitioned into mutually exclusive content area sub-domains. Let $n_j$ represent the number of items in the pool belonging to sub-domain $j$, and $k_j$ represent the number of items belonging to sub-domain $j$ that are administered in each CAT. Note that each $k_j$ is fixed across all CAT administrations. Each CAT is formed by randomly selecting $k_j$ items from sub-domain $j$ for $j = 1, 2, 3, \ldots, J$ (i.e., altogether there are $J$ sub-domains). Therefore, the total fixed test length of each CAT is given by

$$K = \sum_{j=1}^{J} k_j, \tag{A-6}$$

and the total number of items in the pool is given by

$$N = \sum_{j=1}^{J} n_j. \tag{A-7}$$

For a pairwise comparison between any two fixed-length CATs that have been administered, with respect to sub-domain $j$, let $Y_j$ denote the number of shared items. Then, the between-test overlap with respect to sub-domain $j$ for such a pairwise comparison is given by $Y_j/k_j$. Possible values for the random variable $Y_j$ are $y = 0, 1, 2, \ldots, k_j$. From Equation A-2, the expected value of $Y_j$ is

$$\mathrm{E}[Y_j] = \frac{k_j^2}{n_j}. \tag{A-8}$$

If each item in the pool belongs to one and only one sub-domain, then each of the $Y_j$ is independent. If we let the random variable

$$Z = \sum_{j=1}^{J} Y_j \tag{A-9}$$

denote the total number of shared items in a pairwise comparison between any two CATs, then

$$\mathrm{E}[Z] = \sum_{j=1}^{J} \mathrm{E}[Y_j] = \sum_{j=1}^{J} \frac{k_j^2}{n_j}. \tag{A-10}$$

Therefore, under completely randomized item selection, the expected value of the between-test overlap for a fixed-length CAT composed of items from a pool that is partitioned into mutually exclusive content area sub-domains is

$$\mathrm{E}\left[\frac{Z}{\sum_{j=1}^{J} k_j}\right] = \frac{\mathrm{E}[Z]}{\sum_{j=1}^{J} k_j} = \frac{\sum_{j=1}^{J} \frac{k_j^2}{n_j}}{\sum_{j=1}^{J} k_j}. \tag{A-11}$$

**TABLE 1**

**Two Examples (Cases) of Items Administered in Each of 4 CATs**

| CAT | Items administered | | | | |
|---|---|---|---|---|---|
| | | | Case 1 | | |
| $p_1$ | 2 | ⑤ | 4 | 7 | 8 |
| $p_2$ | 1 | ③ | 4 | 6 | 2 |
| $p_3$ | 7 | 9 | 2 | 5 | 8 |
| $p_4$ | 8 | 7 | 1 | 3 | 2 |
| | | | Case 2 | | |
| $p_1$ | 2 | ③ | 4 | 7 | 8 |
| $p_2$ | 1 | ⑤ | 4 | 6 | 2 |
| $p_3$ | 7 | 9 | 2 | 5 | 8 |
| $p_4$ | 8 | 7 | 1 | 3 | 2 |

Case 2 and the circled item numbers are explained on page 14.

**TABLE 2**

**Item Usage Corresponding to Both Cases Presented in Table 1**

| Item ($i$): | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| # of times item was used ($m_i$): 2 | 4 | 2 | 2 | 2 | 1 | 3 | 3 | 1 | 0 | |

**TABLE 3**

**Summary of Empirically Obtained Data From a CAT Simulation Study by Chen and Ankenmann (1999), and Corresponding Algebraically Obtained Data**

| Condition | Empirically obtained data From Chen and Ankenmann (1999) | | | | Algebraically obtained data using Equations 10, 14, and 18 | | |
|---|---|---|---|---|---|---|---|
| | Proportion of Item Pool not Used | $\bar{r}$ | $S_r^2$ | $r_{max}$ | $\overline{T}$ | $\hat{\overline{T}}$ | $\varepsilon(\hat{\overline{T}})$ |
| Without content balancing or item exposure control | .6086 | .0556 | .01376 | 1 | .30316 | .30326 | -.00010 |
| With content balancing | .6306 | .0556 | .01131 | .5660 | .25894 | .25905 | -.00011 |
| With content balancing and item exposure control | .2156 | .0556 | .00229 | .2084 | .09663 | .09676 | -.00013 |
| Random item selection | 0 | .0556 | $7.14 \times 10^{-6}$ | .0629 | .05555 | .05568 | -.00013 |

28

29

## TABLE 4

### Summary of Empirically Obtained Data From a CAT Simulation Study by Chang (1998)

| Condition | $\bar{r}$ | $S_r^2$ | $r_{max}$ | $\bar{T}$ | $S_T^2$ | $T_{max}$ |
|---|---|---|---|---|---|---|
| | | | $k = 30,\ n = 360$ | | | |
| No Control | .08333 | .02419 | 1 | .37110 | .08337 | 1 |
| M&M | .08333 | .02346 | .72012 | .36362 | .08092 | 1 |
| S&H | .08333 | .00171 | .13640 | .10359 | .00652 | .6 |
| D&P | .08333 | .00127 | .12816 | .09877 | .00437 | .4 |
| S&L-U | .08333 | .00169 | .13150 | .10401 | .00656 | .63333 |
| S&L-C | .08333 | .00052 | .10342 | .08947 | .00259 | .33333 |
| | | | $k = 30,\ n = 720$ | | | |
| No Control | .04167 | .01250 | 1 | .34028 | .08043 | 1 |
| M&M | .04167 | .01226 | .64194 | .33629 | .07775 | 1 |
| S&H | .04167 | .00257 | .14348 | .10286 | .01143 | .83333 |
| D&P | .04167 | .00182 | .13536 | .08500 | .00542 | .56667 |
| S&L-U | .04167 | .00256 | .13886 | .10308 | .01152 | .9 |
| S&L-C | .04167 | .00073 | .10032 | .05919 | .00249 | .5 |

*Notes.* The variances of the item exposure rates reported by Chang (1998) were based only on those items in the pool that were administered. The variances, $S_r^2$, reported in this table are adjusted to include all items in the pool and are, as such, estimates.

No Control: No item exposure control was implemented in this condition.
M&M: The 5-4-3-2-1 randomization technique of McBride and Martin (1983).
S&H: The Sympson and Hetter (1985) procedure, with a desired maximum item exposure rate of .10.
D&P: The Davey and Parshall (1995) procedure.
S&L-U: The Stocking and Lewis (1995) unconditional multinomial procedure.
S&L-C: The Stocking and Lewis (1998) conditional procedure.
$S_T^2$: The variance of the between-test overlaps.
$T_{max}$: The maximum value of the between-test overlaps.

**TABLE 5**

**Summary of Between-Test Overlap Data Corresponding to Both Cases Presented in Table 1**

| Pairwise comparison | Between-test overlap | |
|---|---|---|
| | Case 1 | Case 2 |
| $p_1p_2$ | $2/5$ | $2/5$ |
| $p_1p_3$ | $4/5$ | $3/5$ |
| $p_1p_4$ | $3/5$ | $4/5$ |
| $p_2p_3$ | $1/5$ | $2/5$ |
| $p_2p_4$ | $3/5$ | $2/5$ |
| $p_3p_4$ | $3/5$ | $3/5$ |
| Mean ($\bar{T}$) | $16/30$ | $16/30$ |
| Variance ($S_T^2$) | .0356 | .0222 |

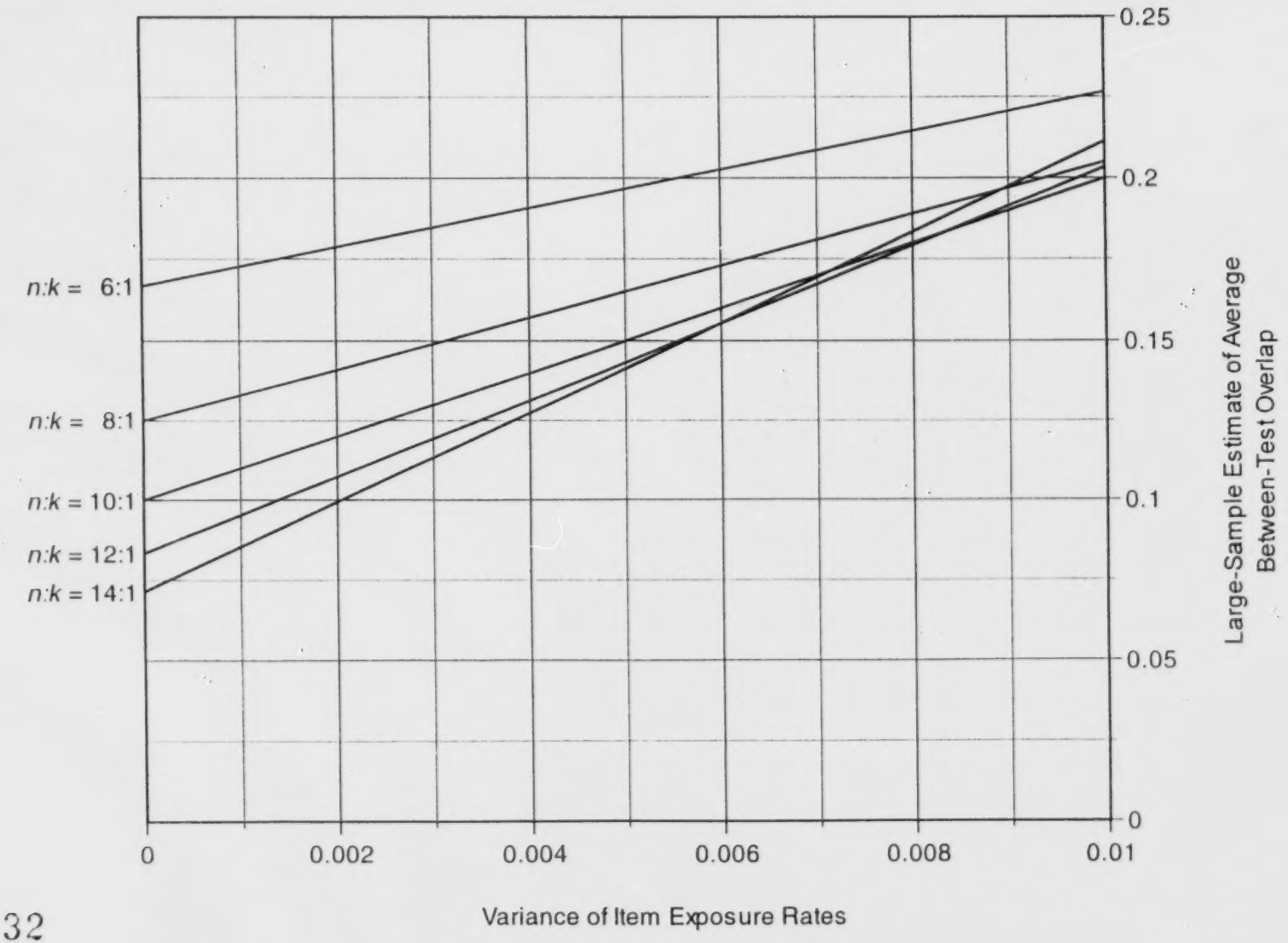


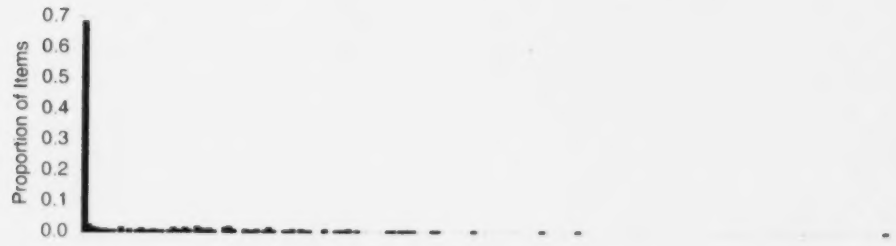




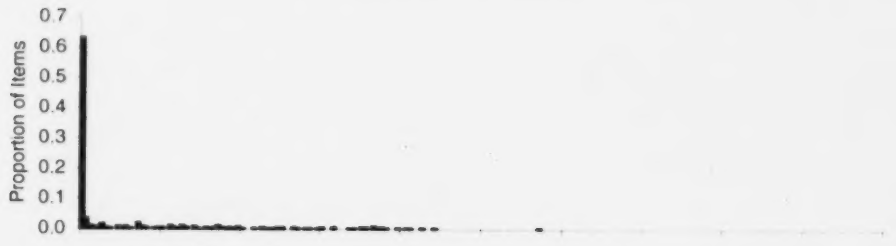

*Figure 1.* A family of curves of the large-sample estimate of average between-test overlaps vs. the variance of the item exposure rates for several pool size to fixed test length ratios

*FIGURE 2:* Frequency distributions of item exposure rates under various CAT simulation conditions studies by Chen and Ankenmann (1999).

(a) Without Content Balancing or Item Exposure Control




(b) With Content Balancing




(c) With Content Balancing and Item Exposure Control

(d) Random Item Selection